




Sanitized clustering against confounding bias

Yinghua Yao^{1,2,3,4}  · Yuangang Pan^{1,2} · Jing Li^{1,2} · Ivor W. Tsang^{1,2,4} · Xin Yao³

Received: 1 June 2023 / Revised: 10 August 2023 / Accepted: 7 October 2023
© The Author(s) 2023

Abstract

Real-world datasets inevitably contain biases that arise from different sources or conditions during data collection. Consequently, such inconsistency itself acts as a confounding factor that disturbs the cluster analysis. Existing methods eliminate the biases by projecting data onto the orthogonal complement of the subspace expanded by the confounding factor before clustering. Therein, the interested clustering factor and the confounding factor are coarsely considered in the raw feature space, where the correlation between the data and the confounding factor is ideally assumed to be linear for convenient solutions. These approaches are thus limited in scope as the data in real applications is usually complex and non-linearly correlated with the confounding factor. This paper presents a new clustering framework named Sanitized Clustering Against confounding Bias, which removes the confounding factor in the semantic latent space of complex data through a non-linear dependence measure. To be specific, we eliminate the bias information in the latent space by minimizing the mutual information between the confounding factor and the latent representation delivered by variational auto-encoder. Meanwhile, a clustering module is introduced to cluster over the purified latent representations. Extensive experiments on complex datasets demonstrate that our SCAB achieves a significant gain in clustering performance by removing the confounding bias.

Keywords Deep clustering · Confounding bias · Mutual information · Non-linear dependence

1 Introduction

Clustering is an essential technique for unsupervised data analysis, whose objective is to partition samples into groups so that the samples in the same group are similar while those from different groups are significantly different (Jain et al., 1999). Standard clustering methods (Cheng, 1995; Modha & Spangler, 2003; Xie et al., 2016) is capable of capturing the desired semantic structure embedded in the clean raw data. However, biases are

Editors: Vu Nguyen, Dani Yogatama.

The first version of this work was done when the first author was at SUSTech.

Extended author information available on the last page of the article

our SCAB can obtain a precise clustering structure in the latent space of complex data robust to the biases. The contributions of this work are summarized as:

- We propose the first deep clustering framework SCAB for clustering complex data contaminated by confounding biases. Unlike existing related studies, SCAB performs semantic clustering in the latent space while minimizing the non-linear dependence between the latent representation and the biases.
- Our theoretical analyses reveal that in our SCAB, (1) the loss for clustering maximizes a lower bound of the mutual information between the data representation and the desired clustering structure; (2) the loss for removing the biases minimizes an upper bound of the mutual information between the data representation and the confounding factor induced by biases.
- We conduct extensive experiments on seven biased image datasets. Empirical results demonstrate the superiority of our sanitized clustering with removing confounding biases, and our SCAB consistently achieves better results than existing baselines.

2 Problem statement and related work

We first introduce standard clustering that neglects the data biases. Then, we motivate our problem setting where data contains confounding biases and discuss the deficiencies of existing work. Last, we compare our setting with two related clustering branches and discuss the issues when their methodologies are applied to our setting.

2.1 Standard clustering

Let $X = [x_1, x_2, \dots, x_N]^T \in \mathbb{R}^{N \times D}$ be a dataset with N samples and D features. Standard clustering is to partition the dataset X into K groups by minimizing inter-cluster similarity and maximizing intra-cluster similarity:

$$\min_{S_x \in \mathbb{S}_{K,x}} F(S_x). \quad (1)$$

$\mathbb{S}_{K,x}$ denotes all feasible K -partitions of X ¹. S_x is a K -partition in raw feature. F is the clustering objective, whose minimization aims at optimizing the quality of clustering. For instance, the k -means clustering objective is $F = \sum_{k=1}^K \sum_{n=1}^N s_{nk} \|x_n - e_k\|_2^2$, where e_k is the k -th cluster centroid. $s_{nk} \in \{0, 1\}$ denotes the cluster assignment which equals 1 if x_n is assigned to the k -th cluster and 0 otherwise.

While classical approaches (Cheng, 1995) conduct clustering in the raw feature space, recent deep clustering methods (Xie et al., 2016; Guo et al., 2017; Huang et al., 2020; Niu et al., 2022) explores clustering-favourable latent representation for a better structure discovery. However, when there are obvious variances resulting from biases present in the data, all standard clustering methods are unavoidably distracted by the confounding biases and the clustering performance will degenerate (see Tables 3, 4).

¹ A K -partition of a set X denotes a collection of K mutually disjoint non-empty subsets whose union is X . Namely, $S_x = (S_1, S_2, \dots, S_K)$, where $\bigcup_{i=1}^K S_i = X$, $S_i \cap S_j = \emptyset$, $1 \leq i \neq j \leq K$.

2.2 Clustering data contaminated by confounding biases

When the data is collected from multiple sources or different conditions, each source may have its own biases. These biases could mask genuine similarities or differences between data points, distorting the desired clustering results (Jacob et al., 2016). In this case, the data source can be said a confounding factor that hinders the accurate clustering structure. In addition, confounding factors that bias the clustering results in other scenarios can also be identified by the domain experts. For instance, in the facial recognition task, whether people wearing glasses or not would impair the recognition results for identity (Sharif et al., 2016).

In order to deliver a precise clustering structure, we consider removing the influence of these confounding biases. We suppose such bias information can be always described by a label indicator, which is an effective encoding for the confounding factor (e.g., a source indicator indicating the data is from source 1, 2, or etc.). Given the complete instance-wise confounding factor, we define our problem setting in the following.

Definition 1 (Sanitized clustering with the removal of confounding bias) Let $X \in \mathbb{R}^{N \times D}$ be a dataset with N samples and D features. Let $C = [c_1, c_2, \dots, c_N]^T \in \{0, 1\}^{N \times G}$ be the corresponding labels with regards to a certain confounding factor c , where $C_{i,j} = 1$ if x_i belongs to class j and $C_{i,j} = 0$ otherwise; G is the number of categories. Our goal is to find a partition $\mathcal{S}_x \in \mathbb{S}_{K,x}$, such that \mathcal{S}_x is uninformative of c . The objective is:

$$\min_{\mathcal{S}_x \in \mathbb{S}_{K,x}} F(\mathcal{S}_x), \quad \text{s.t. } \mathcal{S}_x \perp c, \quad (2)$$

where \perp denotes that two variables are independent.

Existing work. Some work (Jacob et al., 2016; Listgarten et al., 2010; Gagnon-Bartsch & Speed, 2012) targeting the problem (Definition 1) are built on a linear model that assumes the confounding factor is linearly correlated with the data. Mathematically, let $A \in \{0, 1\}^{N \times K}$ denote a group assignment matrix, and each row of $\alpha \in \mathbb{R}^{K \times D}$ denote a cluster centroid. Supposing $C \in \{0, 1\}^{N \times G}$ represents the class matrix converted via the confounding factor c , and each row of $\beta \in \mathbb{R}^{G \times D}$ denotes the centroid of the corresponding category. Then, the linear model is formulated as:

$$X = A\alpha + C\beta + \varepsilon, \quad (3)$$

where ε denotes some prior noise. β can be estimated via a regression model by setting $A\alpha = 0$ (Jacob et al., 2016). By subtracting the bias term $C\beta$, a purified dataset \hat{X} is:

$$\hat{X} = X - C\beta. \quad (4)$$

Then, a regular clustering method like k -means is conducted on \hat{X} to obtain a partition $\mathcal{S}_{\hat{x}}$ (i.e., A and α). Under the linear assumption, the obtained partition thus satisfies the independent constraint, namely, $\mathcal{S}_{\hat{x}} \perp c$.

Deficiencies that make existing approaches impractical for high-dimensional complex data. (1) They are developed in the raw feature space, which is insufficient to discover the underlying structures in terms of the interested factor as well as the confounding factor, i.e., α and β in Eq. (3). (2) Only linear dependence is explored. The removal of the confounding factor is simply via a linear projection, i.e., Eq. (4), which will fail when the data has a non-linear dependence with the confounding factor.

2.3 Related clustering branches

Alternative clustering (Wu et al., 2018) suggests finding an alternative structure w.r.t. the existing clustering result to reveal a new viewpoint of the dataset. Niu et al. (2013); Wu et al. (2019) pursued a novel clustering while minimizing its dependence on the given clustering structure. In particular, the relevance is measured by a specific kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). Given a dataset $X \in \mathbb{R}^{N \times D}$, let $Y = [y_1, y_2, \dots, y_N]^T \in \{0, 1\}^{N \times K_0}$ be an existing clustering result over X , where K_0 is the number of clusters. $y_{ij} = 1$ if x_i belongs to the j -th cluster and $y_{ij} = 0$ otherwise. The aim is to discover an alternative clustering $U \in \mathbb{R}^{N \times K}$ with K clusters on some lower dimensional subspace of dimension $Q (\ll D)$. Let $W \in \mathbb{R}^{D \times Q}$ be a projection matrix. Their objective is usually defined as:

$$\max_{W,U} \text{HSIC}(XW, U) - \lambda \text{HSIC}(XW, Y), \quad \text{s.t.} \quad W^T W = I, U^T U = I. \tag{5}$$

The solution of Eq. (5) can be referred to Niu et al. (2013); Wu et al. (2018).

Alternative clustering vs. our setting (Def. 1). Although starting from a different motivation, Eq. (5) can be a practical implementation form for Eq. (2) by replacing the given clustering structure with the confounding factor. However, obtaining the subspace irrelevant to the confounding factor by a linear projection is not suitable for the high-dimensional complex dataset where the factor is a high-level semantic feature. Meanwhile, such a technique requires storing a full batch of data for clustering, which incurs a heavy memory complexity of $\mathcal{O}(N^2)$.

Fair clustering² that extends group fairness (Feldman et al., 2015) to clustering explores the clustering structure while ensuring a balanced proportion within each cluster regarding some specified sensitive attribute (Chierichetti et al., 2017). With a slight abuse of annotation, suppose the dataset X can be represented as the disjoint union of H protected subgroups in terms of some sensitive attribute a , i.e., $X = \bigsqcup_{h \in [H]} X_h = \bigcup_{h \in [H]} \{(x, h) \mid x \in X_h\}$. For a clustering result $\mathcal{S}_x \in \mathbb{S}_{K,x}$, the balance of each cluster S_k and the whole clustering result \mathcal{S}_x can be respectively defined as:

$$\mathcal{B}(S_k \mid a) = \min_{h \neq h' \in [H]} \frac{|X_h \cap S_k|}{|X_{h'} \cap S_k|} \in [0, 1]; \quad \mathcal{B}(\mathcal{S}_x \mid a) = \min_{k \in [K]} \mathcal{B}(S_k \mid a). \tag{6}$$

The higher the balance of each cluster, the fairer the clustering result will be. A (T, K) -fair clustering (Chierichetti et al., 2017; Kleindessner et al., 2019) is defined as:

$$\min_{\mathcal{S}_x \in \mathbb{S}_{K,x}} F(\mathcal{S}_x), \quad \text{s.t.} \quad \mathcal{B}(\mathcal{S}_x \mid a) \geq T, \tag{7}$$

where T controls the degree of fairness for clustering. Equation (7) pursues a partition where each cluster approximately maintains the same ratio over the sensitive attribute as that in the whole dataset (Chierichetti et al., 2017; Kleindessner et al., 2019).

Fair clustering vs. our setting (Def. 1). Both fair clustering and our problem setting require information about some specific attribute (factor) before conducting clustering. However, fair clustering aims to deliver a clustering structure that meets fairness criteria

² Note that some recent work (Mahabadi & Vakilian, 2020; Vakilian & Yalciner, 2022) which are also called fair clustering are not related to our setting, because they follow the individual fairness (Jung et al., 2019) where group attributes are not specified.

over a certain sensitive attribute. The clustering performance would degrade when imposing such an extra fairness constraint (Chierichetti et al., 2017). In contrast, our target is to improve clustering by eliminating the effect of the confounding factor that distracts the clustering results. Therefore, fair clustering methods (Eq. (7)) cannot be applied to our setting, except a recent deep fair clustering (DFC) (Li et al., 2020). DFC was proposed to learn fair representation for clustering and claimed to adopt stronger fairness criteria than the balance criteria (Eq. (6)). It introduced an adversarial training paradigm in the context of deep standard clustering to encourage clustering structures to be independent of the sensitive attribute. This form of fair clustering objective is the same as ours (Eq. (2)) when the sensitive attribute is designated as the confounding factor. However, the adversarial training increases the difficulty of model training and requires an extra complex constraint to maintain the clustering structure.

3 Sanitized clustering against confounding bias

This section presents a new framework SCAB to deliver desired clustering structures on complex datasets contaminated by confounding biases.

3.1 Deep semantic clustering in the latent space

We perform clustering in the latent space to capture the semantic structure of complex data. Consider a general task (e.g., data reconstruction) that involves encoding the data x into its latent representation z via the posterior $q(z | x)$ (an encoder). The objective of deep semantic clustering includes the objective L for representation learning and the objective F for clustering on the representations (Xie et al., 2016; Boubekki et al., 2021). Namely,

$$\min_{q, \mathcal{S}_z \in \mathbb{S}_{K,z}} L(q, x) + \eta F(\mathcal{S}_z). \quad (8)$$

\mathcal{S}_z denotes a partition in the space where z resides. $\mathbb{S}_{K,z}$ is defined similarly as $\mathbb{S}_{K,x}$ in Eq. (1). η is a trade-off parameter that balances representation learning and clustering.

In particular, we choose Variational AutoEncoder (VAE) (Kingma & Welling, 2014) to compute $L(q, x)$, because VAE includes modeling of $q(z | x)$, and VAE based clustering can obtain good clustering-favorable representations and is effective for various complex datasets (Jiang et al., 2017).

3.2 Clustering on representations invariant to confounding factor

Equation (8) conducts semantic clustering without considering the existence of the confounding bias. To eliminate the negative impact of the bias on the target clustering structure \mathcal{S}_z , we propose deep semantic clustering independent of the confounding factor c . Recalling Eq. (2), our objective is formulated as:

$$\min_{q, \mathcal{S}_z \in \mathbb{S}_{K,z}} L(q, x) + \eta F(\mathcal{S}_z), \quad \text{s.t. } \mathcal{S}_z \perp c. \quad (9)$$

Since a partition \mathcal{S}_z is defined over the whole dataset while c is collected per sample, directly implementing $\mathcal{S}_z \perp c$ is complex and incurs large computational costs. Instead, we

impose an alternative independence constraint between the sample representation z and the confounding factor c , i.e., $z \perp c$, both of which are defined at the sample level.

Proposition 1 *Let \mathcal{Z} be the representation space, and $Z = \{z_1, z_2, \dots, z_N\}^T \in \mathcal{Z}$ be the representation set of the dataset X . Suppose the clustering algorithm \mathcal{A} takes Z as an input and returns a partition \mathcal{S}_z of Z . Namely, $\mathcal{A} : Z \rightarrow \mathcal{S}_z$. If $z \perp c$, then we naturally have $\mathcal{S}_z \perp c$.*

Proposition 1 demonstrates clustering over representations z that is invariant to the confounding factor c can derive a clustering structure \mathcal{S}_z that is uninformative of the confounding factor c . Thus, our objective can be reformulated as:

$$\min_{q, \mathcal{S}_z \in \mathbb{S}_{K_z}} L(q, x) + \eta F(\mathcal{S}_z), \quad \text{s.t. } z \perp c. \tag{10}$$

The independence constraint $z \perp c$ is still a strong condition and is difficult to optimize directly. A natural relaxation of this constraint is to minimize the mutual information $I(z, c)$ (Moyer et al., 2018). Adding the term $I(z, c)$, the objective Eq. (10) becomes:

$$\min_{q, \mathcal{S}_z \in \mathbb{S}_{K_z}} L(q, x) + \eta_1 I(z, c) + \eta_2 F(\mathcal{S}_z). \tag{11}$$

where η_1 and η_2 are the hyper-parameters that balance the three losses. In Eq. (11), the interested clustering factor, which is embedded in the representation z , and the confounding factor c can be semantically described in the latent space (Xie et al., 2016; Vincent et al., 2010). Meanwhile, these two factors are disentangled in the latent space. By optimizing Eq. (11), we can obtain a semantic clustering structure \mathcal{S}_z that is irrelevant to the confounding factor c .

3.3 The overall clustering framework: SCAB

To summarize, our framework jointly trains with three modules. First, the VAE structure is adopted as the feature extractor module for learning semantic features. Further, we introduce one disentangling module over the latent space derived by VAE, to disentangle the confounding factor c and other salient information z encoded in the data (i.e., $z \perp c$). Last, a clustering module based on soft k -means is incorporated within the VAE structure to perform clustering on the factor of interest (embedded in z) only.

3.3.1 Variational autoencoder

Accordingly, we can formulate the statistical (non-linear) dependence between x and c in the latent space, i.e., $p(x, z, c) = p(z, c)p(x | z, c)$ where z is the latent variable of x .

Similar to VAE (Kingma & Welling, 2014), the variational lower bound for the expectation of conditional log-likelihood $\mathbb{E}_{(x,c)} [\log p(x | c)]$ can be deduced as follows:

$$\mathbb{E}_{(x,c)} [\log p(x | c)] \geq \mathbb{E}_{(x,c)} [\mathbb{E}_{z \sim q(z|x)} [\log p(x | z, c)] - KL[q(z | x) || p(z)]]. \tag{12}$$

The conditional decoder $p(x | z, c)$ takes both z and c as input. We simplify the distribution of z to solely depend on the input x , optimized by the encoder $q(z | x)$. $p(z)$ is the prior distribution which is defined as a Gaussian noise.

We parameterize the approximate posterior $q(z | x)$ with an encoder f_ϕ that encodes a data sample x to its latent embedding z , and parameterize the likelihood $p(x | z, c)$ with a

conditional decoder g_θ that produces a data sample conditioned both on the latent embedding z and the observed confounding factor c . Usually, a particle z_n is sampled from $q(z | x)$ for reconstructing x_n (Kingma & Welling, 2014). Then, the loss function (minimization) based on the Monte Carlo estimation of the variational lower bound in Eq. (12) is defined as:

$$\mathcal{L}_{\text{VAE}} = \sum_{n=1}^N \ell_r(x_n, g_\theta(z_n, c_n)) + \sum_{n=1}^N KL[q_\phi(z | x_n) \| p(z)], \quad (13)$$

where ℓ_r denotes the reconstruction loss, which can be instantiated with mean squared loss or cross-entropy loss. \mathcal{L}_{VAE} is used to calculate the first term $L(q, x)$ in Eq. (11).

3.3.2 Disentanglement by minimizing mutual information

By minimizing the mutual information $I(z, c)$ between the latent variable z and the confounding factor c , the bias information is disentangled from other salient information in the latent space.

Lemma 1 (MI upper bound (Moyer et al., 2018)) *The mutual information $I(z, c)$ between the latent representation z and the confounding factor c is subject to an upper bound:*

$$I(z, c) \leq -H(x | c) - \mathbb{E}_{x,c,z \sim q}[\log p(x | z, c)] + \mathbb{E}_x[KL[q(z | x) \| q(z)]]. \quad (14)$$

As $I(z, c)$ is not directly computable, we use its upper bound (Eq. (14)). $H(x | c)$ is a constant and can be ignored. The second term is a reconstruction loss as Eq. (13). The third term on the right of Eq. (14) is intractable to compute and is approximated by its pairwise distances $KL[q(z | x) \| q(z | x')]$ (Moyer et al., 2018):

$$\mathbb{E}_x[KL[q(z | x) \| q(z)]] \approx \sum_x \sum_{x'} KL[q(z | x) \| q(z | x')].$$

The loss function is finally defined as:

$$\mathcal{L}_{\text{MI}} = \sum_{n=1}^N \ell_r(x_n, g_\theta(z_n, c_n)) + \sum_{n=1}^N \sum_{m=1}^N KL[q_\phi(z | x_n) \| q_\phi(z | x'_m)]. \quad (15)$$

The minimization of $I(z, c)$, the second term in Eq. (11), is thus replaced by the minimization of its upper bound, i.e., \mathcal{L}_{MI} .

3.3.3 Clustering over the c -invariant embedding

Eq. (15) helps to filter out the information of the confounding factor c from the latent code z . For the sake of efficiency, we apply k -means algorithm to conduct clustering on the c -invariant embedding z . Particularly, the k -means clustering loss is defined as:

$$\mathcal{L}_{\text{cluster}} = \sum_{n=1}^N \sum_{k=1}^K s_{nk} \|z_n - e_k\|_2^2. \quad (16)$$

$\mathcal{L}_{\text{cluster}}$ is used to compute the third term $F(\mathcal{S}_z)$ in Eq.(11). $\mathbf{e} = \{e_1, e_2, \dots, e_K\}$ are the collection of K centroids. $s_{nk} \in \{0, 1\}$ refers to the group assignment that assigns the latent embedding z to its closest clustering centroid. Namely,

$$\lambda_{nk} = \frac{\exp\left(-\tau\|z_n - e_k\|_2^2\right)}{\sum_{i=1}^K \exp\left(-\tau\|z_n - e_i\|_2^2\right)}, \quad s_{nk} = \begin{cases} 1 & k = \operatorname{argmax}_j \lambda_{nj} \\ 0 & \text{otherwise} \end{cases}, \quad (17)$$

where $k = 1, 2, \dots, K$. τ is the temperature and is set to 5 in the experiment.

Due to the reconstruction loss in VAE (Eq. (13)), the latent representations would contain many sample-specific details, which is detrimental to clustering. We follow (Pan & Tsang, 2021) to introduce the following skip-connection formulation to unify the reconstruction goal and the clustering goal. Namely,

$$\hat{z}_n = h_\psi(z_n, \tilde{z}_n), \text{ where } \tilde{z}_n = \sum_{k=1}^K s_{nk} e_k. \quad (18)$$

Note that \tilde{z}_n is one of K clustering centroids as s_{nk} is a one-hot assignment. h_ψ constructs a new latent representation \hat{z}_n that incorporates not only the original c -invariant embedding z_n but also its belonging clustering centroid \tilde{z}_n as the input of the decoder. h_ψ is implemented as a linear layer.

3.3.4 Objective and optimization of SCAB

Integrating all three modules comes to our new framework Sanitized Clustering Against confounding Bias (SCAB) (Fig. 1). Its final objective is formulated as:

$$\mathcal{L}(\Theta, \mathbf{e}) = \mathcal{L}_{\text{VAE}} + \eta_1 \mathcal{L}_{\text{MI}} + \eta_2 \mathcal{L}_{\text{cluster}}, \quad (19)$$

where $\Theta = \{\theta, \phi, \psi\}$ denote the network parameters and \mathbf{e} represent clustering parameters. η_1 and η_2 are the trade-off parameters.

Clustering structure. After training the model, the clustering structure $\mathcal{S}_z = (S_1, S_2, \dots, S_K)$ is calculated by: $S_k = \{z_n \mid s_{nk} = 1, n = 1, 2, \dots, N\}$, where $k = 1, 2, \dots, K$ and s_{nk} is defined in Eq. (17).

In Eq. (19), two types of parameters, i.e., network parameters Θ , and clustering parameters \mathbf{e} , are coupled together, which hinders them from joint optimization. We adopt coordinate descent to alternatively optimize Θ and \mathbf{e} .

To make our SCAB scalable to large-scale problems, we adopt stochastic gradient updates for all parameters. However, such an update for clustering centroids \mathbf{e} would be unstable because the clustering centroids estimated by different mini-batch data may be of great discrepancy. To overcome this issue, we apply the exponential moving average (EMA) update for the centroids since the EMA update yields good stability (Van Den Oord et al., 2017). Specifically, each centroid e_k is updated online using the assigned neighbor representations in the mini-batches $\{z_b\}_{b=1}^B$:

$$\mu_k^{(t)} := \gamma \mu_k^{(t-1)} + (1 - \gamma) \sum_{b=1}^B s_{bk}^{(t-1)} z_b^{(t-1)}, \quad B_k^{(t)} := \gamma B_k^{(t-1)} + (1 - \gamma) \sum_{b=1}^B s_{bk}^{(t-1)}, \quad e_k^{(t)} := \frac{\mu_k^{(t)}}{B_k^{(t)}}, \quad (20)$$

where $\gamma \in [0, 1]$ is a decay parameter (set to 0.995 by default). t is the iteration index.

3.4 Theoretical analysis

In this section, we theoretically analyze that optimizing network parameters Θ of SCAB in Eq. (19) is equivalent to (1) maximizing the lower bound of the mutual information between the representation and the interested clustering structure, i.e., $\max_z I(z, s)$, while (2) minimizing the upper bound of the mutual information between the representation and the confounding factor, i.e., $\min_z I(z, c)$.

Theorem 2 Assume a fixed clustering structure, i.e., the clustering centroids $\mathbf{e} = \{e_1, e_2, \dots, e_K\}$ and the cluster assignments $\{s_n\}_{n=1}^N$, where s_n is a K -dimensional one-hot vector and s_{nk} is defined in Eq.(17). The minimization of our clustering object $\mathcal{L}_{\text{cluster}}$ is equivalent to maximizing the lower bound of the mutual information between the representation z and the interested clustering structure, represented by the group assignment s , i.e., $I(z, s)$, given the clustering centroids \mathbf{e} .

Proof Based on the definition of mutual information, we have

$$I(z, s) = \int p(z, s) \log \frac{p(z, s)}{p(z)p(s)} dz ds = \int p(z, s) \log \frac{p(s | z)}{p(s)} dz ds.$$

Assume $p(x, c, z, s) = p(x, c)p(z | x, c)p(s | x, c, z) = p(x, c)p(z | x, c)p(s | x, c)$, where $p(s | x, c, z) = p(s | x, c)$ follows the conditional independence. Since $p(s | z) = \int p(x, c, s | z) dx dc = \int \frac{p(z|x,c)p(x,c)}{p(z)} p(s | x, c) dx dc$ is intractable, we introduce an auxiliary distribution $q(s | z)$ as an approximation to $p(s | z)$ (Alemi et al., 2017). Because $\text{KL}[p(s | z) || q(s | z)] \geq 0 \implies \int p(s | z) \log p(s | z) ds \geq \int p(s | z) \log q(s | z) ds$, we obtain

$$\begin{aligned} I(z, s) &\geq \int p(z, s) \log \frac{q(s | z)}{p(s)} dz ds = \int p(z, s) \log q(s | z) dz ds + H(s) \\ &\stackrel{\textcircled{1}}{=} \int p(x, c) p(z | x, c) p(s | x, c) \log q(s | z) dx dc dz ds + H(s) \\ &= \mathbb{E}_{(x,c) \sim p(x,c)} \mathbb{E}_{z \sim p(z|x,c)} \mathbb{E}_{s \sim p(s|x,c)} \log q(s | z) ds + H(s) = L_f(z, s) + H(s). \end{aligned} \quad (21)$$

$\textcircled{1}$ is valid since $p(z, s) = \int p(x, c, z, s) dx dc = \int p(x, c) p(z | x, c) p(s | x, c) dx dc$.

The auxiliary distribution $q(s | z)$ can be naturally defined by our k -means clustering module (Sect. 3.3.3). Accordingly, we have $q(s_{nk} = 1 | z_n) = \frac{\exp(-\tau \|z_n - e_k\|_2^2)}{\sum_{i=1}^K \exp(-\tau \|z_n - e_i\|_2^2)}$. Note that we approximate the posterior $p(z | x, c)$ by the VAE encoder $q(z | x)$ constrained with the minimization of $I(z, c)$ and usually one particle z_n is sampled from $q(z|x)$ to reconstruct x_n (Kingma & Welling, 2014). Together with the given cluster assignment $s_n \sim p(s | x, c)$, we have

$$L_f(z, s) = \sum_{n=1}^N \sum_{k=1}^K s_{nk} \log \frac{\exp(-\tau \|z_n - e_k\|_2^2)}{\sum_{i=1}^K \exp(-\tau \|z_n - e_i\|_2^2)} \stackrel{\textcircled{1}}{\xrightarrow{\tau \rightarrow +\infty}} - \sum_{n=1}^N \sum_{k=1}^K s_{nk} \|z_n - e_k\|_2^2.$$

$\textcircled{1}$ is valid because the value of $q(s_{nk} = 1 | z_n)$ approaches zero for all k except for the one corresponding to the smallest distance (Kulis & Jordan, 2012). Then, we have

$$I(z, s) \geq -\mathcal{L}_{\text{cluster}} + H(s). \quad (22)$$

$H(s)$ can be ignored since it is a constant. We complete the proof. \square

Corollary 1 Fixing the centroids \mathbf{e} as well as the cluster assignments $\{s_n\}_{n=1}^N$, Eq. (19) is subject to the following lower bound:

$$\text{Eq. (19)} \geq -\mathbb{E}_{(x,c)}[\log p(x | c)] + \eta_1 I(z, c) - \eta_2 I(z, s). \quad (23)$$

Because three terms of Eq. (19) are respectively lower bounded according to Eqs. (12), (13), (14), (15), and (22).

From Corollary 1, we conclude that the optimization for Θ given \mathbf{e} is to learn a clustering-favorable representation, which is invariant to the confounding factor c .

Remark 1 (Continuous/Incomplete confounding factor) (1) Our method and theoretical analysis are applicable to the continuous confounding factors as well, as they do not specify the exact form of the confounding factor. We will conduct experiments to demonstrate the efficacy of our SCAB on the continuous confounding factor in Sect. 4. (2) For the known confounding factor without ready-to-use annotations, we additionally collect a small amount of supervision for it to avoid too much manual cost. Then, we can solve the problem in a semi-supervised manner, which will be explored in Sect. 4.4.

4 Experiments

Dataset. We conduct experiments on six image datasets (*UCI-Face*, *Rotated Fashion*, *MNIST-USPS*, *Office-31*, *CIFAR10-C*, *Rotated Fashion-Con*) and one signal-vector dataset (*HAR*) containing confounding factors that would bias the clustering results (see Table 1). In particular, *Rotated Fashion* is constructed by introducing the rotation factor into the *Fashion-MNIST* dataset. Specifically, we pick up images from cloth categories, i.e., “T-shirt/top”, “Trouser”, “Pullover”, “Dress”, “Coat” and “Shirt”, for simplicity. We first randomly sample 1,000 images from each of the six classes (zero degree). Then, each image is augmented with four views of 72, 144, 216, and 288 degrees, respectively. For *Office-31*, we select samples from Amazon and Webcam as training data following Li et al. (2020). *Rotated Fashion-Con* is constructed similarly, but the rotation angle is set to a continuous range of 0 to 60 degrees. For *CIFAR10-C*, we consider one in each main category of corruptions, namely, frost, Gaussian blur, impulse noise, and elastic transform for simplicity.

Implementations. We employ the AE architecture described in Xie et al. (2016) for all datasets. The encoder is a fully connected multi-layer perceptron (MLP) with dimensions D -500-500-2000- d . D is the dimension of input. d is the dimension of centroids, which is set to 10 for all datasets. All layers use ReLU activation except the last. The decoder is mirrored of the encoder. Compared with those AE-based clustering methods (Xie et al., 2016; Guo et al., 2017), our SCAB introduces only one extra linear layer for Eq.(18), which bring negligible network parameter overhead. We apply SCAB to raw data for *UCI-Face*, *Rotated Fashion*, *MNIST-USPS*, *HAR* and *Rotated Fashion-Con* considering their simplicity. Inspired by the recent state-of-art (SOTA) clustering methods (Tsai et al., 2021; Niu et al., 2022), which rely on structured representations to achieve superior performance on

Table 1 Statistics of datasets. K denotes the number of clusters. G denotes the number of categories or range of the values

Dataset	#Sample	#Dim	K	Confounding factor (G)
<i>UCI-Face</i> (Bay et al., 2000)	1,872	32×30	4	identity (20)
<i>Rotated Fashion</i> (Xiao et al., 2017)	30,000	28×28	5	cloth category (6)
<i>MNIST-USPS</i> (Lecun et al., 1998; Hull, 1994)	67,291	32×32	10	source of digit (2)
<i>Office-31</i> (Saenko et al., 2010)	3,612	$224 \times 224 \times 3$	31	domain source (2)
<i>CIFAR10-C</i> (Hendrycks & Dietterich, 2019)	40,000	$32 \times 32 \times 3$	10	corruption type (4)
<i>HAR</i> (Anguita et al., 2013)	10,299	561	6	subject (30)
<i>Rotated Fashion-Con</i> (Xiao et al., 2017)	30,000	28×28	6	rotation angle (0–60)

complex datasets, we apply SCAB to the extracted features for *Office-31* and *CIFAR10-C* considering their complexity. We use ImageNet-pretrained ResNet50 to extract features for Office-31 following the SOTA clustering method on Office-31 (Li et al., 2020). We use MoCo (He et al., 2020) to extract features for *CIFAR10-C* following the SOTA clustering method on *CIFAR10-C* (Niu et al., 2022). Note that these feature extractors do not utilize any supervision regarding the datasets. We adopt the Adam optimizer. The default learning rate, training epoch, and batch size are $5e-4$, 1, 000, and 256, respectively.

Baselines. The method that removes the confounding factor in the raw space via linear projection, i.e., RUV (Jacob et al., 2016) (Eq.(3), Eq.(4)), is included as our first baseline. Further, we extend RUV to eliminate the confounding factor in the latent space. In Particular, we first train AE to obtain the latent representations for *UCI-Face*, *Rotated Fashion*, *MNIST-USPS* and *HAR*. We use the extracted features described above as the representations for *Office-31* and *CIFAR10-C*. Then, we apply RUV to remove the bias information from the representations. We name these two baselines as RUV_x and RUV_z , respectively. We also consider Iterative Spectral Method (ISM) (Wu et al., 2019) and Deep Fair Clustering (DFC) (Li et al., 2020) as our baselines since these two methods can be deemed as the same objective as ours (Eq.(2)). We do not compare with other fair clustering methods since they have different goals from our setting (see Sect. 2.3). *For a fair comparison, we take raw images of UCI-Face, Rotated Fashion and MNIST-USPS and extracted features of Office-31 and CIFAR10-C as input for all the baselines except for RUV_x , which takes raw data as input.* ISM, DFC and RUV are designed for the discrete confounding factor and cannot be applied to the continuous one, so they are not run on *Rotated Fashion-Con*.

Metrics. We evaluate different clustering methods with two widely-used clustering metrics, i.e., accuracy (ACC), normalized mutual information (NMI) and Adjusted Rand Index (ARI). For both two metrics, values range between 0 and 1, and a higher value indicates better performance.

4.1 Performance comparison

Quantitative results of our SCAB and various baselines that can remove the confounding factor are summarized in Table 2. It shows that: (1) **SCAB obtains superior results on all datasets.** This is because it adopts an effective non-linear dependence measure and a joint training paradigm, which can learn clustering-favorable representations invariant to the confounding factor. (2) SCAB can be applied for removing the continuous confounding factor (see *Rotated Fashion-Con* in Table 3) while existing baselines cannot. (3) **Latent**

Table 2 SCAB compared with baselines that can remove the confounding factor w.r.t. ACC (\uparrow), NMI (\uparrow) and ARI (\uparrow). The best results are highlighted in bold. The second-best results are underlined

Dataset	Metric	ISM	DFC	RUV _x	RUV _z	SCAB
<i>UCI-Faces</i>	ACC	0.763	0.394	0.380	0.539	0.824
	NMI	0.454	0.087	0.163	0.322	0.570
	ARI	0.482	0.054	0.042	0.198	0.554
<i>Rotated Fashion</i>	ACC	N.A	0.539	0.579	0.993	<u>0.985</u>
	NMI	N.A	0.351	0.516	0.969	<u>0.940</u>
	ARI	N.A	0.248	0.318	0.982	<u>0.961</u>
<i>MNIST-USPS</i>	ACC	N.A	<u>0.825</u>	0.457	0.785	0.919
	NMI	N.A	<u>0.789</u>	0.379	0.756	0.837
	ARI	N.A	–	0.236	0.690	0.831
<i>Office-31</i>	ACC	0.659	<u>0.692</u>	0.186	0.673	0.724
	NMI	0.671	<u>0.718</u>	0.232	0.714	0.728
	ARI	0.495	–	0.065	0.548	0.565
<i>CIFAR10-C</i>	ACC	N.A	0.283	0.208	<u>0.357</u>	0.458
	NMI	N.A	0.186	0.085	0.317	<u>0.311</u>
	ARI	N.A	0.105	0.040	<u>0.087</u>	0.274
<i>HAR</i>	ACC	0.556	0.722	<u>0.732</u>	0.715	0.823
	NMI	0.477	0.632	0.689	<u>0.791</u>	0.830
	ARI	0.368	0.546	0.598	<u>0.671</u>	0.754

Table 3 SCAB compared with standard clustering w.r.t. ACC (\uparrow), NMI (\uparrow) and ARI (\uparrow) on four simple image datasets and one signal-vector dataset

Dataset	Metric	<i>k</i> -means	IDEC	SCAB
<i>UCI-Faces</i>	ACC	0.266	0.356	0.824
	NMI	0.002	0.069	0.570
	ARI	−0.001	0.058	0.554
<i>Rotated Fashion</i>	ACC	0.487	0.602	0.985
	NMI	0.414	0.611	0.940
	ARI	0.260	0.465	0.961
<i>MNIST-USPS</i>	ACC	0.506	0.789	0.919
	NMI	0.447	0.766	0.837
	ARI	0.333	0.689	0.831
<i>HAR</i>	ACC	0.600	0.680	0.823
	NMI	0.589	0.733	0.830
	ARI	0.461	0.632	0.754
<i>Rotated Fashion-Con</i>	ACC	0.369	0.387	0.576
	NMI	0.228	0.287	0.399
	ARI	0.139	0.191	0.329

space is better than raw space. Non-linear correlation is better than linear correlation. RUV_z achieves better performance than RUV_x, which shows that removing the confounding factor in the latent space is more effective than that in the raw space. RUV_z obtains worse results than our SCAB on four datasets since RUV_z simply adopts linear projection and heavily relies on the extracted representations beforehand, which cannot deal with these complex datasets where the desired clustering factor and the confounding factor

are coupled non-trivially in the latent space. (4) **DFC originally designed for two categories degenerates on the dataset with more categories** (i.e., *UCI-Faces*, *Rotated Fashion*, and *CIFAR10-C*). On one hand, more categories may increase the difficulty of adversarial training, making it unable to effectively remove the confounding factor. On the other hand, the constraint requires training a DEC (Xie et al., 2016) for each category of data. For example, it needs to train a DEC on around 93 images for *UCI-Face*, which would suffer from insufficient training samples. (5) **ISM cannot be executed on large-scale datasets**, i.e., *Rotated Fashion*, *MNIST-USPS* and *CIFAR10-C*. ISM requires a memory complexity of $\mathcal{O}(n^2)$ and needs to store a data matrix with a size larger than $10k \times 10k$ for these datasets, which is beyond our computing capacity.

4.2 Efficacy of removing the confounding factor for clustering

To demonstrate the gain of clustering that takes into account the removal of the confounding factor, we include the comparison with standard clustering methods – k -means (Bishop, 2006), IDEC (Guo et al., 2017),³ PICA (Huang et al., 2020) and SPICE (Niu et al., 2022)⁴ in Tables 3 and 4. We apply PICA and SPICE only on *Office-31* and *CIFAR10-C* considering that they were proposed for complex image datasets. For a fair comparison, we take raw images of *UCI-Face*, *Rotated Fashion*, *MNIST-USPS*, *HAR* and *Rotated Fashion-Con* and extracted features of *Office-31* and *CIFAR10-C* as input for the methods except for PICA. PICA takes raw images of all datasets as input since it needs to conduct image augmentations for partition confidence maximization (Huang et al., 2020).

Improved by removing the confounding factor Tables 3 and 4 show that: compared with standard clustering methods, our SCAB achieves superior performance on all datasets. It verifies the claim that our SCAB which explicitly removes the influence of the confounding factor performs better than the standard clustering methods. Note that PICA obtains poor results since it conducts clustering on raw features (k -means on MoCo extracted feature achieves better results than PICA on raw features also reported in Tsai et al. (2021)). And SPICE performs worse than IDEC because it applies a discriminative model for clustering, which is more vulnerable to the confounding factor than IDEC which is AE-based clustering.

Invariant representations

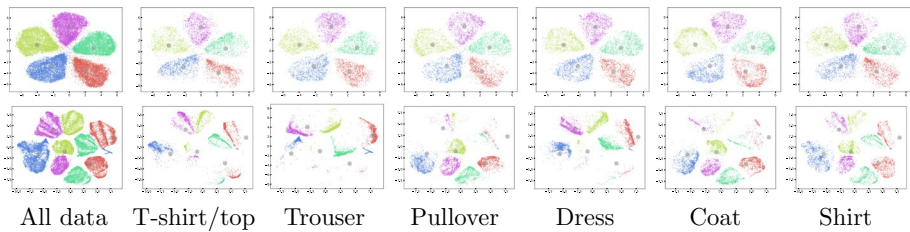
To further illustrate the effectiveness of removing the confounding factor, we visualize the latent representations and the clustering centroids for our SCAB and IDEC (i.e., standard clustering that ignores the confounding factor) on *Rotated Fashion*, respectively. From the t-SNE visualization of our SCAB (the first row of Fig. 2), we can see that: (1) the clusters are well separated and the centroids are located at the center of each cluster. (2) These categories' representations are not only well aligned with each other, but also the whole data's representations. This demonstrates that our SCAB's latent representations are invariant to the confounding factor, i.e., the cloth category label. (3) Each centroid represents one of the five rotation angles in the dataset. In addition, the reconstruction of the centroids

³ IDEC is a representative AE-based clustering method.

⁴ PICA and SPICE are recently proposed self-supervised clustering methods. SPICE is the SOTA method.

Table 4 SCAB compared with standard clustering w.r.t. ACC (\uparrow), NMI (\uparrow) and ARI (\uparrow) on two complex image datasets

Dataset	Metric	k -means	IDEC	PICA	SPICE	SCAB
<i>Office-31</i>	ACC	0.648	0.634	0.440	0.231	0.724
	NMI	0.689	0.690	0.536	0.341	0.728
	ARI	0.506	0.500	0.305	0.117	0.565
<i>CIFAR10-C</i>	ACC	0.247	0.420	0.220	0.313	0.458
	NMI	0.225	0.380	0.178	0.294	0.311
	ARI	0.074	0.257	0.082	0.149	0.274

**Fig. 2** t-SNE on latent representations and clustering centroids from SCAB (1st row) and IDEC (2nd row) on *Rotated Fashion*, respectively. The big grey dots are the centroids. The small dots are the representations, of which the colors denote the ground truth category labels

is exactly the Fashion-MNIST objects, which demonstrates our SCAB captures semantic clustering structures.

The t-SNE visualization of IDEC (the second row of Fig. 2) shows that: (1) IDEC obtains an inferior clustering structure due to the negative impact of the confounding factor. Specifically, the cloth category introduces variances into the data, making the derived structure away from the desired one w.r.t. the rotation factor. (2) These categories' representations are neither aligned with each other nor with the representation of the entire data. It demonstrates that IDEC's latent representations are corrupted by the confounding factor, i.e., variances of cloth category.

Disentangled centroid reconstruction

We can reconstruct the centroids conditioned on the confounding factor for SCAB. Figure 3 shows that (1) the latent embedding z and the confounding factor c are well disentangled. In particular, the information of the confounding factor is well captured by c . (2) The centroids can capture clear structures, i.e., rotation angles for *Rotated Fashion*, the pose angle for *UCI-Face*, and the digit type for *MNIST-USPS*, respectively. On *Office-31* and *CIFAR10-C*, we do not reconstruct the centroids on these datasets as the extracted features are used as model input.

Figure 4 shows that (1) IDEC does not have the ability to disentangle the confounding factor c from the latent space. (2) Its centroids do not capture all rotation angles in the dataset as they are distracted by the cloth categories. For example, e_1 and e_2 represent the shirt and the trouser with the same angle, respectively.

4.3 Ablation study

We study the effectiveness of each module by excluding it from our SCAB (Fig. 1).

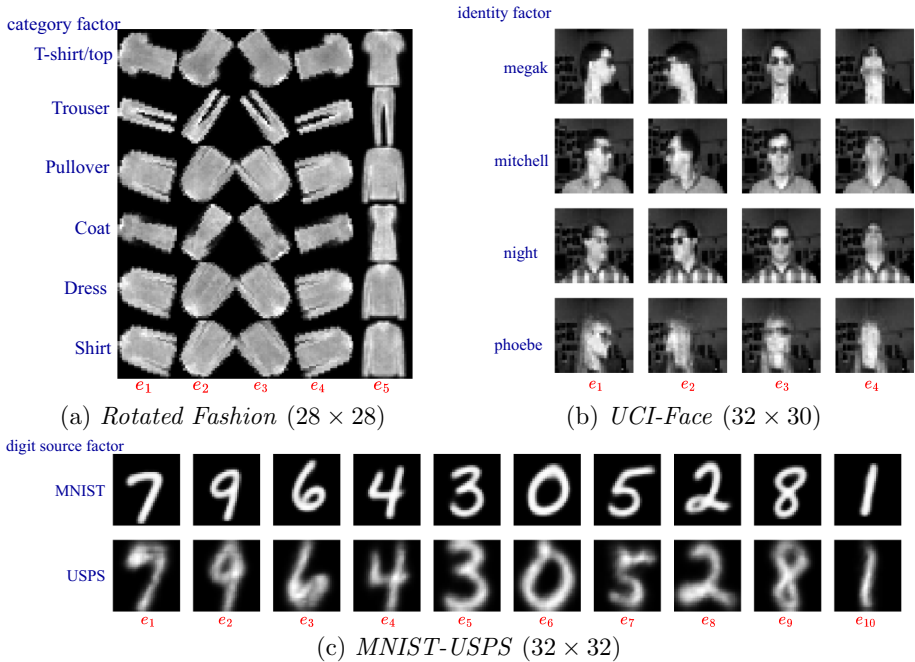


Fig. 3 Centroids' reconstruction of SCAB on *Rotated Fashion*, *UCI-Face* and *MNIST-USPS*, respectively. Each column is conditioned on the same clustering centroid. Each row is conditioned on different labels of the cloth category factor, the identity factor, and the digit source factor, respectively

Fig. 4 Centroids' reconstruction of IDEC on *Rotated Fashion*

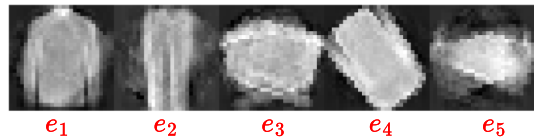


Table 5 shows that: (1) our SCAB gets the best results, which justifies the necessity of each module. (2) Without the disentanglement module to remove the confounding factor via mutual information, the clustering performance drops significantly since the confounding factor would distract desired the clustering results. (3) A poor clustering structure is obtained without the clustering module because it fails to derive clustering-friendly representations. (4) The clustering performance is worse when excluding both the clustering module and the disentanglement module.

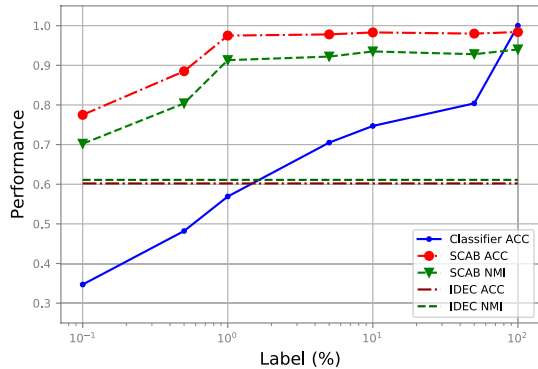
4.4 Extension to the incomplete confounding factor

We explore the performance of SCAB given different amounts of labeled data w.r.t. the confounding factor on *Rotated Fashion*. Applying SCAB to this semi-supervised setting, we first train a classifier on the labeled data and use it to predict labels for the remaining

Table 5 Ablation study of SCAB on *Rotated Fashion*. “Clu” means the clustering module. “Dis” means the disentanglement module

Metric	w/o Clu	w/o Dis	w/o Clu & Dis	SCAB
ACC	0.513	0.857	0.487	0.985
NMI	0.376	0.803	0.414	0.940
ARI	0.277	0.757	0.260	0.961

Fig. 5 Clustering performance of SCAB given partial labels w.r.t. the confounding factor on *Rotated Fashion*. “Classifier ACC” is the test accuracy of the classifier. x axis is the ratio of labeled data



unlabeled data. Then SCAB is naturally applied to these fully-labeled data. Particularly, we employ a convolutional neural network classifier for the classification. IDEC is adopted as the baseline without removing the confounding factor following the same setting as SCAB.

We plot the test accuracy of the classifier (calculated on the remaining unlabeled data) and the clustering performance (ACC and NMI) of SCAB in Fig. 5 with the percentage of labeled data from 0.1 to 100%. It shows that (1) compared to IDEC which ignores the confounding factor, our SCAB can improve the clustering performance even with a very small amount of labeled data. (2) When there are less than 0.5% labeled data, the test accuracy of the classifier is low, smaller than 0.5. Accordingly, the results of SCAB are relatively not so good since there are more than 50% samples assigned with wrong labels. (3) When the labeled data is larger than 1%, there are more than 50% samples assigned with true labels. Though the percentage of label noise is still very high, SCAB can perform well since the correct labels dominate and the structured representations can be robust to label noise. In conclusion, our SCAB can work well even given a small amount of labeled data regarding the confounding factor.

5 Conclusion

We have introduced a general framework SCAB for a new stream of clustering that aims to deliver clustering results invariant to the pre-designated confounding factor. SCAB is the first deep clustering framework that can eliminate the confounding factor in the semantic latent space of complex data via a non-linear dependence measure with theoretical guarantees. We have demonstrated the efficacy of SCAB on various datasets using label indicators of the confounding factor. In the future, we can extend our SCAB to more types of data, e.g., text/ time series data. In addition, while this study focuses on sanitized clustering given the known confounding factor with (partially) labeled supervision, it is interesting to

explore clustering with unindicated confounding factors. Last, theoretical analysis on the confounding factor that is not fully observed is also a potential direction.

Author Contributions Idea: YY; Methodology (including literature review): YY, YP, JL; Experiment: YY; Writing - original draft: YY; Writing - comments/edits: all; Supervision: IWT and XY.

Funding This work was supported in part by the A*STAR Centre for Frontier AI Research; in part by the AISG Grand Challenge in AI for Materials Discovery (Grant No. AISG2-GC-2023-010); in part by the A*STAR C222812019; in part by the A*STAR Pitchfest for ECR 232D800027; in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07X386); and in part by the Program for Guangdong Provincial Key Laboratory (Grant No. 2020B121201001).

Data availability All datasets used in this work are available online and clearly cited.

Code availability The code of this work is available at <https://github.com/EvaFlower/SCAB>.

Declarations

Conflict of interest The authors have no financial or non-financial interests to disclose that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publishing Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


- Alemi A. A., Fischer I., Dillon J. V., et al. (2017). Deep variational information bottleneck. In: ICLR
- Anguita, D., Ghio, A., Oneto, L., et al. (2013). A public domain dataset for human activity recognition using smartphones. *21th European symposium on artificial neural networks* (pp. 437–442). CIACO: Computational Intelligence and Machine Learning (ESANN).
- Bay, S. D., Kibler, D. F., Pazzani, M. J., et al. (2000). The UCI KDD archive of large data sets for data mining research and experimentation. *ACM SIGKDD Explorations Newsletter*, 2(2), 81–85.
- Benito, M., Parker, J., Du, Q., et al. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1), 105–114.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*, (Vol. 4). Springer.
- Boubekki, A., Kampffmeyer, M., Brefeld, U., et al. (2021). Joint optimization of an autoencoder for clustering and embedding. *Machine Learning*, 110(7), 1901–1937.
- Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8), 790–799.
- Chierichetti, F., Kumar, R., Lattanzi, S., et al. (2017). Fair clustering through fairlets. In: *NeurIPS*, 30, 5029–5037.
- Feldman, M., Friedler, S. A., Moeller, J., et al. (2015). Certifying and removing disparate impact. *SIGKDD*, 10, 259–268.
- Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3), 539–552.

- Guo, X., Gao, L., Liu, X., et al. (2017). Improved deep embedded clustering with local structure preservation. *IJCAI*, 17, 1753–1759.
- He K., Fan H., Wu Y., et al. (2020) Momentum contrast for unsupervised visual representation learning. In: CVPR, pp 9729–9738
- Hendrycks, D., Dietterich, T. G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR
- Huang, J., Gong, S., Zhu, X. (2020). Deep semantic clustering by partition confidence maximisation. In: CVPR, pp 8846–8855
- Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5), 550–554.
- Jacob, L., Gagnon-Bartsch, J. A., & Speed, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1), 16–28.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.
- Jiang Z., Zheng Y., Tan H, et al (2017) Variational deep embedding: an unsupervised and generative approach to clustering. In: IJCAI, pp 1965–1972
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 118–127.
- Jung C, Kannan S, Lutz N (2019) A center in your neighborhood: Fairness in facility location. arXiv preprint [arXiv:1908.09041](https://arxiv.org/abs/1908.09041)
- Kingma, D. P., Welling M. (2014) Auto-encoding variational bayes. In: ICLR
- Kleindessner, M., Samadi, S., Awasthi, P., et al. (2019). Guarantees for spectral clustering with fairness constraints. In: ICML, PMLR, pp 3458–3467
- Kulis, B., Jordan, M. I. (2012) Revisiting k-means: new algorithms via bayesian nonparametrics. In: ICML, pp 1131–1138
- Lecun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Li, P., Zhao, H., Liu, H. (2020). Deep fair clustering for visual learning. In: CVPR, pp 9070–9079
- Listgarten, J., Kadie, C., Schadt, E. E., et al. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38), 16465–16470.
- Mahabadi, S., Vakilian, A. (2020). Individual fairness for k-clustering. In: ICML, PMLR, pp 6586–6596
- Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine Learning*, 52, 217–237.
- Moyer, D., Gao, S., Brekelmans, R., et al. (2018). Invariant representations without adversarial training. In: NeurIPS, pp. 9102–9111
- Niu, D., Dy, J. G., & Jordan, M. I. (2013). Iterative discovery of multiple alternative clustering views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1340–1353.
- Niu, C., Shan, H., & Wang, G. (2022). Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31, 7264–7278.
- Pan, Y., Tsang, I. (2021). Streamlining em into auto-encoder networks. In: OpenReview
- Saenko, K., Kulis, B., Fritz, M., et al. (2010). Adapting visual category models to new domains. In: ECCV, pp 213–226
- Sharif, M., Bhagavatula, S., Bauer, L., et al. (2016) Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: ACM Conference on Computer and Communications Security, pp. 1528–1540
- Tsai, T. W., Li, C., Zhu, J. (2021). Mice: Mixture of contrastive experts for unsupervised image clustering. In: ICLR
- Vakilian, A., Yalciner, M. (2022) Improved approximation algorithms for individually fair clustering. In: AIST-ATS, PMLR, pp. 8758–8779
- Van Den Oord, A., Vinyals, O., Kavukcuoglu, K. (2017). Neural discrete representation learning. *NeurIPS* pp. 6309–6318
- Vincent, P., Larochelle, H., Lajoie, I., et al. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12), 201.
- Wu, C., Ioannidis, S., Sznajder, M., et al. (2018) Iterative spectral method for alternative clustering. In: AIST-ATS, pp 115–123
- Wu, C., Miller, J., Chang, Y., et al. (2019). Solving interpretable kernel dimensionality reduction. *NeurIPS* pp 7915–7925
- Wu, S., Yuksekgonul, M., Zhang, L., et al. (2023) Discover and cure: Concept-aware mitigation of spurious correlation. In: ICML
- Xiao, H., Rasul, K., Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint [arXiv:1708.07747](https://arxiv.org/abs/1708.07747)

Xie, J., Girshick, R., Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In: ICML, pp 478–487

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yinghua Yao^{1,2,3,4}  · Yuangang Pan^{1,2} · Jing Li^{1,2} · Ivor W. Tsang^{1,2,4} · Xin Yao³

✉ Ivor W. Tsang
ivor.tsang@gmail.com

✉ Xin Yao
xiny@sustech.edu.cn

Yinghua Yao
eva.yh.yao@gmail.com

Yuangang Pan
yuangang.pan@gmail.com

Jing Li
j.lee9383@gmail.com

¹ CFAR, Agency for Science, Technology, and Research (A*STAR), Singapore 138632, Singapore

² IHPC, Agency for Science, Technology, and Research (A*STAR), Singapore 138632, Singapore

³ Computer Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen 518055, China

⁴ Australian Artificial Intelligence Institute, University of Technology Sydney (UTS), Sydney 2007, Australia